Computer Science and Application 计算机科学与应用, 2025, 15(4), 225-235
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154094

Hans汉斯

基于语义无向带权图的文本零水印算法

李　波，刘　微*

沈阳理工大学信息科学与工程学院，辽宁 沈阳

收稿日期：2025年3月11日；录用日期：2025年4月11日；发布日期：2025年4月18日

摘　要

在生成式语言模型兴起的今天，人工智能为文本创作和传播带来了前所未有的变革，但是生成式语言模型的广泛应用也带来了版权保护的问题。本研究基于文本的语义特征，提出了一种创新的文本零水印算法，通过语义相似度编码模型将文本的基础粒度编码为高维向量，接着利用文本粒度的高维语义嵌入向量的方向各异性，构建文本语义特征图，对文本特征进行相关性分析实现相似度的评估。经实验证明，本文所提出的零水印算法，在误判率方面的表现较好；在鲁棒性上，对同义改写和文本添加攻击具有良好的抵抗力，对文本的删除攻击具有一定的鲁棒性。

关键词

文本相似度，文本零水印，版权保护

Text Zero-Watermarking Algorithm Based on Semantic Undirected Weighted Graph

Bo Li, Wei Liu*

School of Information Science and Engineering, Shenyang Ligong University, Shenyang Liaoning

Received: Mar. 11th, 2025; accepted: Apr. 11th, 2025; published: Apr. 18th, 2025

Abstract

With the rise of generative language models, artificial intelligence has brought unprecedented changes to text creation and dissemination, but the widespread application of generative language models has also brought the problem of copyright protection. Based on the semantic features of the text, this study proposes an innovative text zero watermark algorithm, which encodes the basic granularity of the text into high-dimensional vectors through the semantic similarity coding model, and then uses the directional heterogeneity of the high-dimensional semantic embedding vectors

*通讯作者。

文章引用: 李波, 刘微. 基于语义无向带权图的文本零水印算法[J]. 计算机科学与应用, 2025, 15(4): 225-235.
DOI: 10.12677/csa.2025.154094

of the text granularity to construct a text semantic feature map, and analyzes the relevance of the text features to achieve similarity evaluation. Experiments show that the zero-watermark algorithm proposed in this paper has a better performance in terms of false positive rate. In terms of robustness, it has good resistance to synonymous rewriting and text addition attacks, and has a certain robustness to text deletion attacks.

Keywords

Text Similarity, Text Zero-Watermarking, Copyright Protection

1. 引言

随着信息技术的飞速发展，尤其是互联网和社交媒体平台的普及，文本内容的传播变得更加广泛且难以追踪。传统的版权保护手段如数字版权管理(DRM)和显式水印已经表现出一定的局限性，尤其是在文本类内容的保护上。这些传统方法要么可能影响文本的可读性，要么容易被恶意用户篡改或去除。相比之下，文本零水印技术以其不干扰原文、难以察觉的特点，成为一种理想的版权保护解决方案。随着自然语言处理技术的不断进步，文本零水印的研究和应用前景广阔，未来它将成为数字版权保护和信息安全领域的重要工具。

文本零水印(Text Zero-watermarking)技术是一种新兴的信息隐藏方法，它通过在文本特征中嵌入隐形标识来保护文本版权或进行数据验证。与传统的文本水印技术不同，零水印技术无需对原始文本内容做出修改，避免了对文本可读性的影响。零水印在版权保护、数字内容认证等领域具有重要的应用价值。

本文将探讨文本零水印技术的基本原理、发展现状及其面临的挑战。本文简要回顾文本水印的历史和现有技术的不足之处。接着介绍所提出零水印方法的基本概念及其实现方式，分析其在确保版权保护、数据隐私及内容真实性方面的优势。

2. 相关研究以及相关技术

2.1. 文本零水印研究现状

当前的文本水印技术主要分为两种，一种是以改变文本内容的嵌入式文本水印，另一种是不改变文本的不可见零水印。

零水印在当前又主要分为两类，主要是通用型文本零水印以及基于文本体裁的文本零水印。前者主要有龚礼春[1]提出的医疗文本零水印，该文本零水印主要是基于实体命名识别技术，抽取医疗文本中的实体特征，将文本的实体特征构建文本零水印，基于特征的脆弱性实现对文本篡改识别。但是该零水印算法仅仅只能识别文本的数据完整性，不能识别文本的篡改程度。张娜等人[2]提出的零水印方法主要是基于文本的主题词和文本语句的信息熵来构建零水印，主题词由 TF-IDF 抽取并通过同义词词林进行编码方便计算相似度，接着逐句计算全文语句的信息熵，并构建全文信息熵的统计特征。该方法在一定程度上具有良好的鲁棒性，但是零水印的构建过于依赖 TF-IDF 抽取的主题词，对于难负例的效果不如预期。戴夏菁等人[3]提出了一种基于 word2vec 的文本零水印算法，该算法主要是提取文本中的中频次词语，然后基于词嵌入对中频词进行编码，并利用 SVD(奇异值分解)技术对高维向量降维。但是该算法的

总体鲁棒性不够，且中频词对于文本的特征唯一性表征不足。

基于文本体裁的文本零水印对于特定类型的文本具有良好的鲁棒性。姚然[4]基于说明文的文本特征，提取说明文的事物、写作顺序和副词的重要程度。胡毅光[5]基于记叙文的特征，提取文本的中心句、状中结构和顺承句。上述基于特定体裁的文本零水印需要依赖文本的特定特征，提取特征的算法需要精心设计，复杂度较高，零水印特征在第三方信任机构注册所需的存储空间较大且对通用的文本效果不好。

基于上述相关研究工作，本文提出一种基于文本语义特征的零水印，旨在对文本的语义攻击具有良好的鲁棒性。

2.2. 相关技术

2.2.1. 对比学习

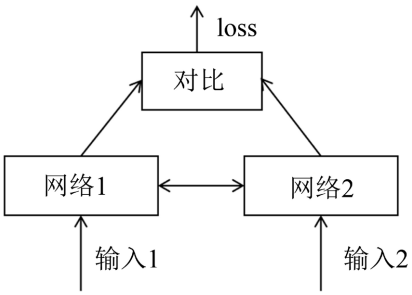

Figure 1. Diagram of the siamese neural network
图 1. 孪生神经网络图

对比学习的一种典型结构为孪生网络(Siamese Network)，如图 1 所示，网络 1 和网络 2 结构完全一样，而且他们的参数相同。当输入 1 和输入 2 分别输入到网络 1 和网络 2 时，样本将会转化为同一高维空间的向量，通过对比两者嵌入向量之间的相似度，通常是余弦相似度或者欧氏距离，来计算网络的损失函数，然后通过反向传播来更新网络参数。对比学习的一个重要用途为拉近正样本在高维空间的距离，将负样本在高维空间的距离变大。

2.2.2. Transformer Encoder

Transformer Encoder 是基于自注意力机制构建的深度神经网络架构核心组件，其通过堆叠多个结构相同的编码层实现序列数据的特征抽象与上下文建模。每个编码层包含多头自注意力子层和前馈神经网络子层，其中多头自注意力机制通过并行计算多组查询–键–值映射关系，捕获序列内部元素间的长程依赖与动态交互模式，前馈神经网络则借助非线性变换提升特征表达能力。各子层均采用残差连接与层归一化操作，有效缓解梯度消失问题并加速模型收敛。通过引入位置编码矩阵，该架构克服了传统循环神经网络无法并行计算的缺陷，同时保留了序列元素的相对位置信息。在多种自然语言处理任务中 BERT [6]由 Transformer Encoder 通过多层级联的特征提取，生成具有丰富语义表征的上下文向量，可以为下游任务提供相应适合的语义嵌入向量。

2.2.3. 余弦相似度

余弦相似度是一种基于高维向量空间模型的相似性度量方法，通过计算两个嵌入向量在方向上的夹角余弦值来评估其相似程度，范围为−1 到 1 之间，相似性越高，向量两者之间的余弦相似度越大。其核心思想在于将数据对象映射至高维空间，通过向量间夹角的几何特性表征相似性，而对向量的绝对模长具有不变性。两个语句向量的余弦相似度计算公式如下所示：

$$\text{CosineSimilarity}(A,\ B)=\frac{A\cdot B}{\|A\|\times\|B\|}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}} \tag{1}$$

2.2.4. 皮尔逊相关性分析

皮尔逊相关系数是统计学中衡量两个变量间线性相关程度的标准化指标，在特定情况对矩阵相似度分析有很好的效果[7]，能够反映变量协同变化的趋势特征。其本质是通过数据中心化处理后计算余弦相似度，消除量纲影响的同时捕捉线性关联模式。取值范围是−1 到 1 之间，值越大表示线性相关性越强。对于两组数据 X 和 Y 的皮尔逊相关性分析，公式如下所示：

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\times\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}} \tag{2}$$

3. 文本零水印算法设计

本文基于 BERT 作为特征提取器构建一种文本零水印，直接以文本的语义作为构建零水印的基础。首先对文本进行预处理，接着利用 BERT 作为文本的编码器为每一个文本粒度编码构建基础粒度向量，然后基于余弦相似度构建文本的无向概率转移图，将图的权值边转化为邻接矩阵，大大减少了零水印的所需储存空间。当遇到争议文本时，对文本进行同样的文本零水印构建操作，接着比对两文本之间的零水印相似度，当达到所设阈值时即可判定文本是否为同一文本。算法的整体流程图如图 2 所示：

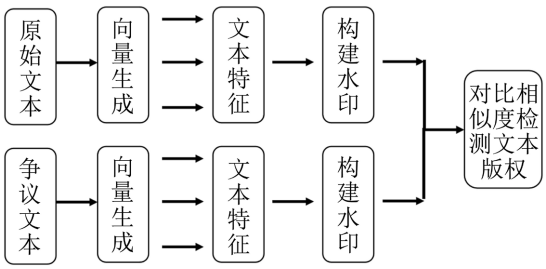

Figure 2. Diagram of the overall flow chart of the text zero-watermarking algorithm
图 2. 文本零水印算法整体流程图

3.1. 文本零水印构建

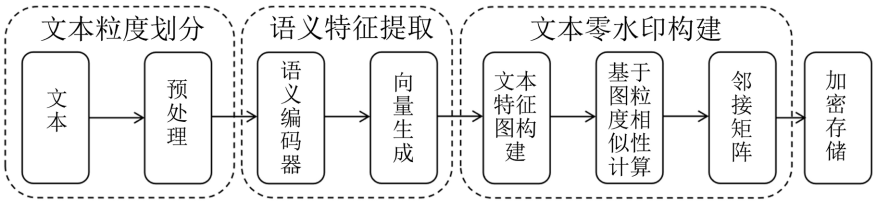

Figure 3. Flowchart of the construction algorithm with text zero-watermarking
图 3. 文本零水印构建算法流程图

本文所提出的文本零水印构建算法如图 3 所示，水印构造详细步骤如下：

1) 对文本进行粒度划分。划分粒度作为构建文本特征的基础，主要分为两种划分方式，当文本段落数小于 4 时，我们将对文本进行语句划分，即以文本的每一个句子作为输入语义编码器的基础粒度；当段落数量大于等于 4 时，我们将对文本进行段落划分，文本的每一个段落直接作为语义编码器的基础粒度。

2) 语义编码器的构建以及训练。本文采用 BERT 作为预训练模型，通过在 LCQMC 数据集[8]上进行微调，使得模型在语义判断上具有良好的效果。通过采用对比学习框架，拉近相似句子对的余弦距离。具体模型构建如图 4 所示，在模型的最底层，将两个句子为一组输入到 BERT 模型中，每一个句子前加一个 cls 标签作为整句语义的表示，通过大量 Transformer Encoder 双向编码后，在 BERT 的输出端输出一个[n + 1, 768]维度的张量，其中 n 表示该句子的 token 数，即连同 CLS 标签，每一个 token 都会被编码为 768 维的向量，如图[c, T1,⋯, Tn]表示。接着通过平均池化[9]将整个句子的 token 向量转化为 768 维的向量代表整句语义，最后通过匹配层对比两个句子语义向量的余弦相似度，经 SoftMax 分类层判断两个句子是否具有相同的语义。

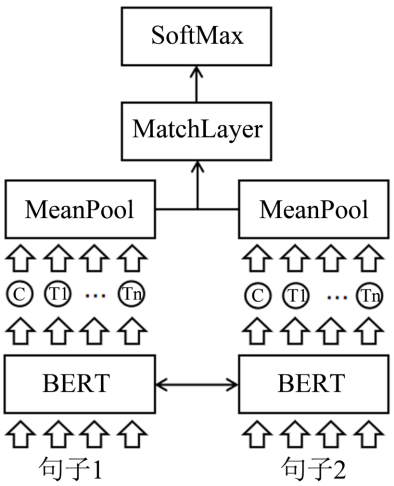

Figure 4. Semantic encoder training network diagram
图 4. 语义编码器训练网络图

3) 文本的无向带权图构建。将文本每一个粒度输入到语义编码器后，文本转化为 n 个高维向量，此时向量作为文本的特征具有较好的准确性，但是由于向量的维度太高，作为零水印存储在第三方数据库中需要大量的存储空间，因此本文提出了一种基于文本无向带权图的特征构建方法，保留文本特征同时大大降低了文本零水印所需的存储空间。具体如图 5 所示：

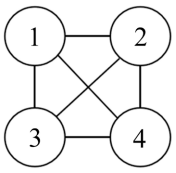

Figure 5. Text undirected weighted maps
图 5. 文本无向带权图

在上图中我们以 4 个基础粒度的文本为例构建文本无向带权图，其中数字代表文本的粒度顺序，如文本句子上下顺序或者文本段落的顺序，通过语义步骤 2 中的语义编码模型将文本的基础粒度转化为 768 维度的向量，无向图的边由两个粒度语义嵌入向量之间的余弦相似度构建，因此 4 粒度的文本可以构建一个 6 条边的无向带权图。一个文本图的定义式为 G = (V, E)。其中图节点 V 为基础粒度经编码器生成的语义向量，无向边 E 为每两个不同基础粒度的余弦相似度，如 Sim (v1, v2)表示第一条无向边。最后将无向边权值储存到邻接矩阵中作为文本的零水印特征。

4) 对于所提取的特征进行 RSA 加密，交由第三方可信任机构存储，进一步增强零水印的安全性。

3.2. 文本零水印检测

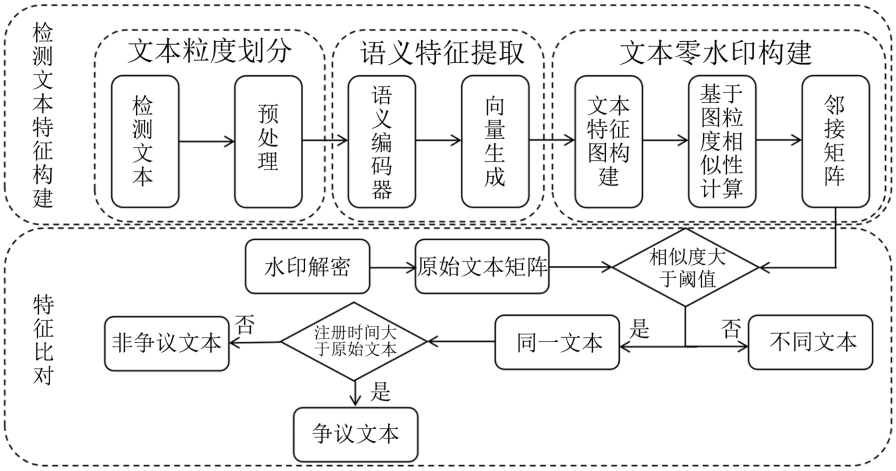

Figure 6. Flowchart of text zero-watermarking detection
图 6. 文本零水印检测流程图

当出现具有争议文本情况时，文本的检测总体流程图如图 6 所示，具体步骤如下：

1) 首先对待检测文本进行文本零水印构建，构建步骤同 3.1 所描述相同。

2) 对第三方机构的原始文本零水印进行解密，获取原始文本的邻接矩阵。

3) 特征相似度检测，本文对两个邻接矩阵相对应元素进行皮尔逊相关性分析，其原理是基于语义相似度高的句子在向量空间中具有很强的同向性，即两者的语义嵌入向量趋近于一个重合的方向，和其他所有不相似的句子的余弦相似度分布应具有接近的分布。直观表示如下图 7 所示：

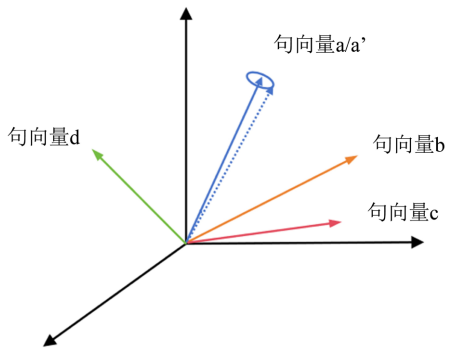

Figure 7. Textual semantic vector distribution feature map
图 7. 文本语义向量分布特征图

在上图的语义空间中，句向量 a'通常表示文本收到攻击之后的句嵌入向量，但是由于攻击者在盗取文本同时必须保证句子的语义相同或者相似，因此通过编码器后句向量 a 和 a'具有很小的夹角。同理，对于其他的语句也是如此，所有的语义嵌入向量都保证在一个很小的夹角范围内“转动”，形成一个很小的锥形结构，文本的语义向量分布呈现一个整体稳定的特征。因此，在基于文本特征图构建的邻接矩阵中，所有的矩阵元素都是在一定范围内上下变动，对两个邻接矩阵相对应元素进行皮尔逊相关性分析是可行的。

4) 最后比对文本零水印注册时间，若注册时间大于原始文本零水印注册时间，则证明所检测文本和原始文本有版权争议，有可能构成抄袭。

4. 实验及结果分析

4.1. 实验环境

本文的实验环境以及相关参数如下表 1 所示：

Table 1. Experimental environmental parameters
表 1. 实验环境参数

实验环境	参数
操作系统	Windows 11
内存	16GB
CPU	Intel i9-12900H
GPU	RTX 3060
Python	3.9
Pytorch	2.2.2 + cu121

BERT 语义编码器训练相关参数如下表 2 所示：

Table 2. Semantic encoder training hyperparameters
表 2. 语义编码器训练超参数

超参数	参数值
Learning Rate	2e-5
Epoch	3
Max_length	512
Hidden_size	768
Batch_size	64
Optimizer	AdamW

4.2. 数据集以及评估指标

1) 用于语义训练的数据集 LCQMC，该数据集为二分类数据集，每一个样本格式为标签 0 或者 1 的句子对，共有训练数据 238,766 条，正样本 138,574 条，负样本 100,192 条。

2) 用于测试文本零水印的数据集，本文为确保随机性，随机取得 20 篇文本，随机选择两两不同的文本作为负样本对，对文本进行不同程度的攻击构建文本正样本对。

3) 评估指标主要有用于评估语义模型性能的 F1 值和准确率 Acc (Accuracy)值，两者值越大证明模型的语义分类效果越好。用于评估文本零水印性能的皮尔逊相关系数，值越大证明文本对之间越相似，可以直接用作零水印鲁棒性评估标准。

4.3. 语义模型性能分析

4.3.1. LCQMC 数据集训练结果

在数据集 LCQMC 训练后，测试结果如下表 3 所示：

Table 3. BERT test results on the dataset LCQMC
表 3. BERT 在数据集 LCQMC 上测试结果

F1	Acc
90.3	90.4

在测试数据集上 F1 值达到了 90.3，Acc 值达到了 90.4，证明模型对语义的判断具有很好的效果。

4.3.2. 相似语句对余弦相似度展示

Table 4. Example of semantic embedding vector similarity
表 4. 语义嵌入向量相似度示例

句子 1	句子 2	余弦相似度
昨晚的暴雨过后，清晨的空气格外湿润清新。	昨夜的暴雨过后，清晨的空气显得格外湿润。	0.9115
这个世界，真的如它展现的那样简单吗？	这个世界，真的简单吗？	0.8493
小山整把济南围了个圈儿，只有北边缺着点口儿。	这就是济南的冬天	0.2130

在表 4 中可以看出相似度较高的句子具有更大的余弦相似度，而不同的句子的余弦相似度较低。

4.4. 文本零水印性能分析

4.4.1. 零水印算法误判率分析

在误判率分析实验中，本文采用不同的文本对来对文本零水印进行评估，就不同的文本对来说，零水印的相似度越低，证明零水印在误判率方面具有更好的效果。

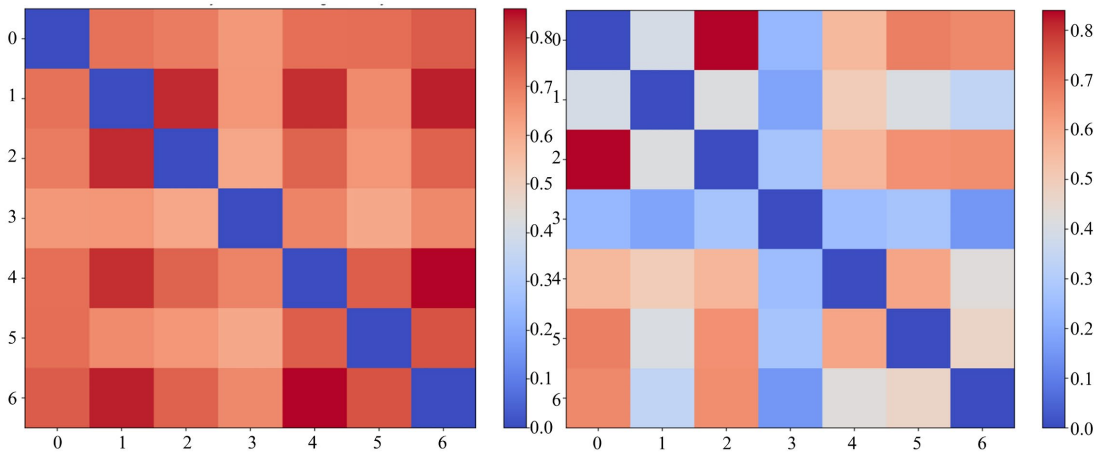

Figure 8. Textual semantic vector similarity distribution feature map
图 8. 文本语义向量相似度分布特征图

以《合欢树》和《秋天的怀念》两文为例，各个粒度之间的转移热力图如上图 8 所示，可以看出不同粒度之间的相似度分布完全由文本的语义决定。在对比相似度时，由于对角线表示每一个粒度的自传相似度，其数值在每一个文本中都是相同的，且下三角元素和上三角元素为关于对角线对称，因此本文在计算相似度时只考虑矩阵的上三角部分。

本文随机选择了 15 个不同文本对误判率进行分析，文本对皮尔逊相关系数结果如下图 9 所示：

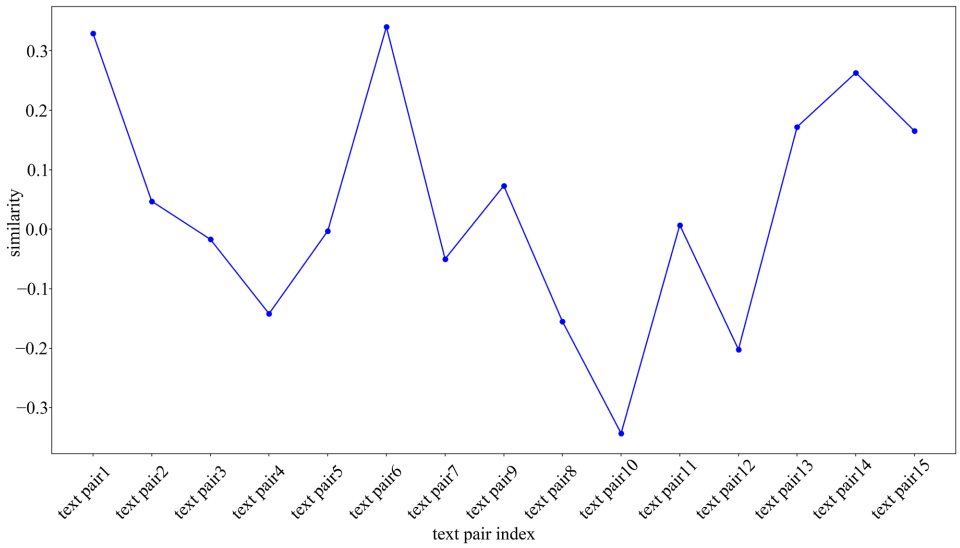

Figure 9. Line charts of different text similarities
图 9. 不同文本相似度折线图

由上图可以看出，本文提出的零水印算法具有很高的文本区分度，相似度最高仅只有 0.34，大部分的不同文本对相似度低于 0.1，因此可以判定该算法误判率较低。

4.4.2. 零水印算法鲁棒性分析

1) 文本语义攻击，本文将传统的文本的同义词替换和句式变换攻击统一，模仿攻击者在不改变语义的情况下对文本进行同义改写。对文本的句子以每 10%的比例对文本进行同义改写，对改写后文本与原始文本邻接矩阵的皮尔逊相关系数平均值作为相似度结果，结果如下图 10 所示：

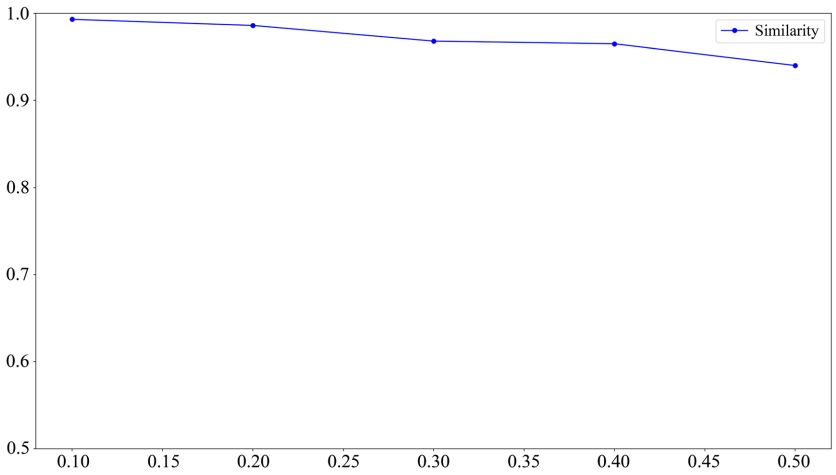

Figure 10. Zero watermark similarity at different semantic attack ratios
图 10. 不同语义攻击比例下的零水印相似度

由上图可看出，零水印算法对文本的语义攻击具有极强的鲁棒性，在同义改写攻击比例达到 50%时，零水印仍然具有 0.94 的相似度。

2) 文本句子删减攻击，在保证文本的主体结构相同的情况下，对冗余或者非重要句子进行随机删除，同样是以每 10%作为梯度，平均值结果如下图 11 所示：

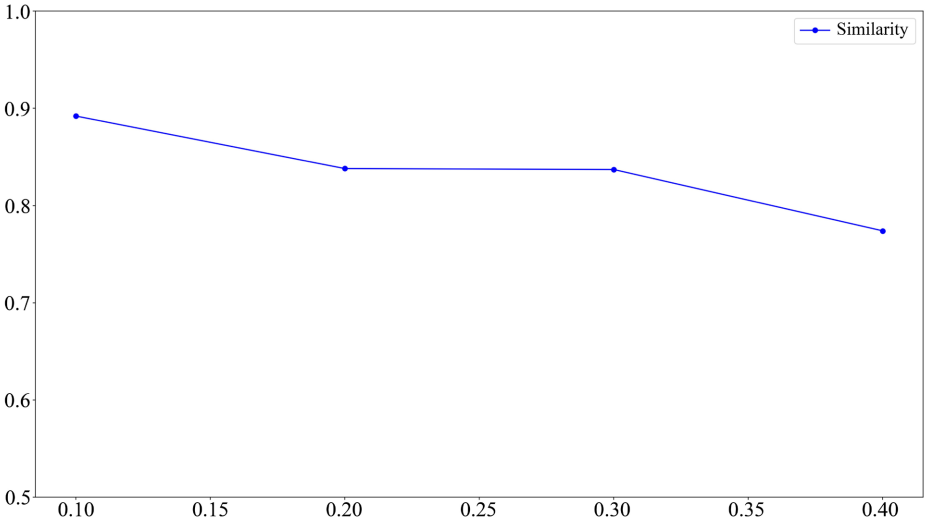

Figure 11. Zero watermark similarity at different text deletion attack ratios
图 11. 不同文本删除攻击比例下的零水印相似度

由上图可以看出，对于文本的删除攻击，必然会导致语义的改变，但是冗余或者非重要句子的删除并不会导致语义发生太大的变化，因此在 20%~30%区间，零水印的整体相似度相对稳定，当删除攻击比例达到 40%时，零水印的相似度开始逐渐下降，但是仍然具有 0.77 的相似度，远远超过不同文本对之间的相似度。

文本句子增加攻击，在保证文本的主体结构相同的情况下，对文本随机添加无关以及重复句子，同样是以每 10%作为梯度，平均值结果如下图 12 所示：

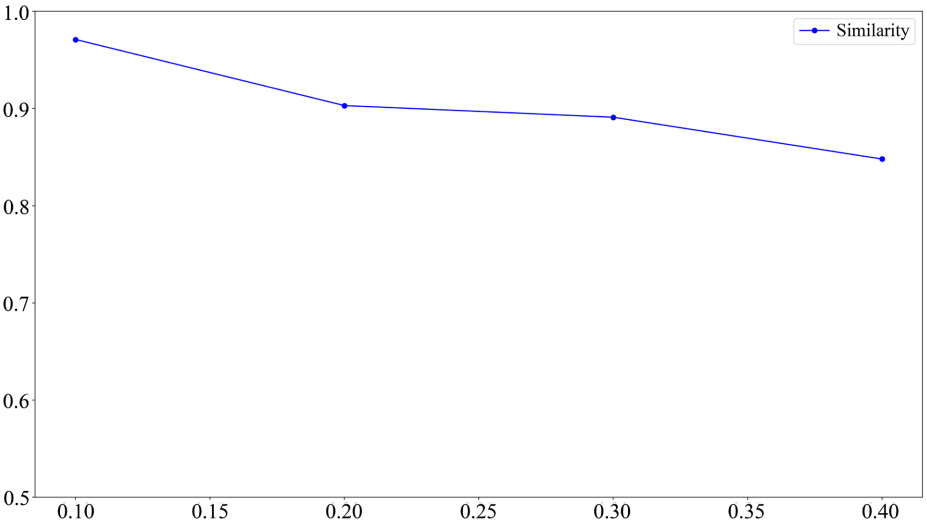

Figure 12. Zero watermark similarity at different text addition attack ratios
图 12. 不同文本添加攻击比例下的零水印相似度

由上图可以看出，对于文本的句子添加攻击，同样会导致语义的改变，但是相对于文本的删除攻击，零水印的整体相似度相对稳定，当句子添加攻击比例达到 40%时，零水印的相似度具有 0.85 的相似度，因此可以判断该零水印算法对句子添加攻击具有较好的鲁棒性。

5. 总结

本文基于文本语义相似度模型提出了一种新的零水印算法，根据文本自身语义结构构建了文本语义特征图，接着由文本粒度之间的无向带权图生成邻接矩阵作为零水印，最后由皮尔逊相关性分析对比文本对之间的相似度。通过实验表明，本文提出的零水印算法具有较强的抵抗同义改写攻击的能力，在针对文本添加方面具有较强的鲁棒性，且在对文本删减攻击上具有一定的鲁棒性。本文提出的零水印算法在对文本版权保护方面，具有一定的实用价值。

参考文献

[1] 龚礼春, 姚晔, 唐观根, 等. 基于命名实体识别的医疗文本零水印方案[J]. 密码学报, 2020, 7(5): 643-654.

[2] 张娜, 张琨, 张先国, 等. 基于主题词与信息熵编码的文本零水印算法[J]. 计算机与数字工程, 2021, 49(8): 1612-1618.

[3] 戴夏菁, 徐谊程, 王馨娅, 等. 基于 Word2Vec 的中文文本零水印算法[J]. 软件工程, 2023, 26(1): 19-23.

[4] 姚然. 说明文零水印算法研究与设计[D]: [硕士学位论文]. 兰州: 兰州大学, 2022.

[5] 胡毅光. 记叙文零水印算法研究与设计[D]: [硕士学位论文]. 兰州: 兰州大学, 2024.

[6] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 4171-4186.

[7] 旷怡, 邓家俊, 段斌. 基于模糊认知图的学生学习效果预测方法[J/OL]. 东南大学学报(自然科学版), 1-11. http://kns.cnki.net/kcms/detail/32.1178.N.20250214.1646.002.html, 2025-03-06.

[8] Liu, X., Chen, Q., Deng, C., *et al*. (2018) Lcqmc: A Large-Scale Chinese Question Matching Corpus. *Proceedings of the* 27*th International Conference on Computational Linguistics*, Santa Fe, 20-26 August 2018, 1952-1962.

[9] Reimers, N. and Gurevych, I. (2019) Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks.